Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

ARTICLE TYPE

Bayesian Inference of Dependent Kappa for Binary Ratings

Ananda Sen*[1,2] | Pin Li[1,4] | Wen Ye[1] | Alfred Franzblau[3]

[1]Department of Biostatistics, University of Michigan, Michigan, U.S.A
[2]Department of Family Medicine, University of Michigan, Michigan, U.S.A
[3]Department of Environmental Health Sciences, University of Michigan, Michigan, U.S.A
[4]Department of Public Health Science, Henry Ford Health System, Michigan, U.S.A

Correspondence
*Ananda Sen, 1018 Fuller St. Ann Arbor, Michigan 48104-1213 Office: (734) 998-0334 Email: anandas@med.umich.edu

Summary

In medical and social science research, reliability of testing methods measured through inter- and intra-observer agreement is critical in disease diagnosis. Often comparison of agreement across multiple testing methods is sought in situations where testing is carried out on the same experimental units rendering the outcomes to be correlated. In this paper, we first developed a Bayesian method for comparing dependent agreement measures under a grouped data setting. Simulation studies showed that the proposed methodology outperforms the competing methods in terms of power, while maintaining a decent type I error rate. We further developed a Bayesian joint model for comparing dependent agreement measures adjusting for subject and rater level heterogeneity. Simulation studies indicate that our model outperforms a competing method that is used in this context. The developed methodology was implemented on a key measure on a dichotomous rating scale from a study with six raters evaluating three classification methods for chest radiographs for pneumoconiosis developed by the International Labor Office.

KEYWORDS:
Correlated Kappa; Test of homogeneity; Bayesian inference; Grouped data; Covariate adjustment

1 | INTRODUCTION

In medical and social science research, analyzing inter- or intra-observer agreement provides a useful means of assessing the reliability of a rating system. Such assessments are particularly relevant in the field of radiology, biomarker research, and survey research, among others. Agreement can be sought across different measurement systems, or across repeated evaluations using the same systems. Such congruence or reliability plays a critical role in disease diagnostics and prognosis. High values of agreement indicate consensus in diagnosis and interchangeability of measuring techniques.

When measurements are made on a binary scale, Cohen's kappa coefficient[1] serves as the most widely employed measure to assess agreement for categorical outcomes. Kappa easily extends to nominal measurement scales with more than two categories. For ordinal scales, weighted versions of kappa are typically preferred[2,3]. In this paper, we will focus our attention on measurements with binary scales, assuming each of N subjects to be rated by two raters. Let p_{ab} be the probability of being assigned to category a by the first rater and to category b by the second rater ($a, b = 0, 1$). The kappa coefficient is

$$\kappa = \frac{P_o - P_e}{1 - P_e} = \frac{2(p_{11} - p_{1.}p_{.1})}{p_{1.} + p_{.1} - 2p_{1.}p_{.1}} \tag{1}$$

where $p_{1.} = p_{10} + p_{11}$ and $p_{.1} = p_{01} + p_{11}$. Here $P_o = p_{00} + p_{11}$ is the observed proportion of agreement and $P_e = p_{1.}p_{.1} + (1 - p_{1.})(1 - p_{.1})$ is the expected proportion of agreement on the basis of chance alone. In the dichotomous setting, kappa has an attractive interpretation. When the raters are interchangeable, i.e. $p_{1.} = p_{.1}$, kappa matches the correlation coefficient between the two rating scales. Although kappa can technically vary between -1 and 1, negative values are typically rare. Landis and Koch[4] have qualitatively assigned ranges of kappa values to varying degrees of agreement. Inference for a single kappa measure is detailed in the books by Fleiss et al.[5] and Shoukri[6]. Several authors have investigated extensions of kappa to multi-rater versions[7] and weighted version for multi-category polytomous response scales[8]. In diagnostic testing, often the interest rests on comparison of competing markers or testing methods. Procedures for comparing two or more independent kappa statistics have been considered in the literature[9,10]. Regression models of kappa incorporating subject level heterogeneity have been studied by Lipsitz et al.[11]. A comprehensive review of early work on kappa and associated agreement measures appears in Banerjee et al.[12].

In situations where comparison of agreement is naturally conducted using the same set of subjects read by a group of raters perhaps using different test methods, the distribution of these agreement measures for different test methods are correlated. Kappa statistics generated in these situations are henceforth referred to as dependent kappas. This is in contrast to comparing kappa statistics derived from independent subgroups of subjects such as different age-groups, or males vs. females. Donner et al.[13] relaxed the assumption of independence adopted in Donner et al.[9] and developed model-based procedures for testing the equality of two dependent kappa statistics for measurements with binary scales. Mckenzie et al.[14], Vanbelle and Albert[15] used resampling methods to compare correlated kappa statistics. By contrast, Barnhart and Williamson[16] proposed a weighted least-squares approach. Kang et al.[17] considered inference on kappa statistics when the binary responses are naturally clustered such as those obtained from a physician-patient diad, where each physician manages multiple patients. Yang and Zhou[18] developed inference for weighted kappa statistic also under a similar framework but in the case of ordinal responses.

Kappa, in its original form, is an aggregate measure and hence does not directly account for subject-level heterogeneity. However, kappa may depend on covariates, i.e. characteristics of the raters or the subjects being rated. Lipsitz et al[11] developed a subject-level kappa as a function of covariates. Regression models of these probabilities implicitly define a covariate-adjusted kappa. Nelson and Edwards[19] introduced a model-based agreement measure different from kappa that is amenable to covariate adjustment. Ma et al.[20] focused on inference for kappas arising from a longitudinal study in presence of missing data.

Our goal is to develop a set of models and methodology for comparing dependent kappa and for identifying potential factors influencing agreement in the presence of a hierarchical structure of data. Donner et al.'s[13] method that directly addresses the comparison of dependent kappas was based on the assumption of exchangeability of raters, and the performance of their method in situations where exchangeability does not hold is unknown. We specifically focus on the Bayesian methodology, which enjoys several attractive advantages over the non-Bayesian methods. Statistical inference using the non-Bayesian methods are predicated on large-sample normality of the estimated kappa. The asymptotic standard error is complex[21] and does not readily lend itself to an expression for comparing correlated kappa's. The Bayesian method, on the other hand, is a finite-sample strategy and is applicable to small-sample situations. Further, a Bayesian framework promotes an expert system where subject-matter knowledge can be readily incorporated and updated as part of the process.

In this paper, we explore our methodology under two different data schemes. First, we address the situation where only grouped data of counts is available for assessment of agreement. This is often the case when secondary data is made available only at the aggregate level in order to mask personal health identifiers. Basu et al.[10] developed a Bayesian method in the context of analyzing grouped data for testing the homogeneity of kappa coefficient across different independent samples. In Section 3, we extend Basu et al.'s method under a Bayesian framework to deal with the problem of comparing correlated kappa statistics computed over the same sample of subjects when only summary data are available. Subsequently, under the assumption that complete rating data are available at the individual level, we adapt and modify a regression model proposed by Lipsitz et al.[11] that accounts for the effect of rater and patient level characteristics. Lipsitz et al.'s model used a two-stage approach, ignoring the uncertainty in the plugged-in estimates of case-specific marginal probabilities carried over from the first stage, thereby inducing potential bias. Williamson et al.[22] proposed a similar method using two sets of generalized estimating equations to model kappa for measuring dependent categorical agreement data, which is subject to the same caveat as in Lipsitz et al.[11]. By contrast, our method jointly estimates parameters in the models from the two stages in a Bayesian setting, thereby properly accounting for the uncertainty. We consider only binary rating scales in this paper. In Section 5, we indicate how the models and methods can potentially extend to multi-category agreement data under dependence.

The remainder of the article is organized as follows. The example that motivated this study is presented in Section 2. Section 3 introduces a Bayesian model for grouped data to compare dependent kappa statistics. Detailed simulation results are presented comparing the proposed methodology to competing approaches. In Section 4, a model-based approach is presented that offers a joint analysis with subject-level data under a Bayesian framework. Our methodology is implemented on a key measure obtained from an example in radiology that serves as the motivating example. Finally, we conclude our paper in Section 5 with some directions for future research. Appendix A contains the relevant codes, with additional tables and figures included in Appendix

B. This study was approved by the University of Michigan Institutional Review Board (IRBMED approval #2002-0855), under which only fully de-identified data will be made available upon request.

2 | A MOTIVATING EXAMPLE

Work in this article was originally motivated by the need to evaluate a standardized classification method for chest radiographs for pneumoconiosis developed by the International Labour Office (ILO) based in Geneva, Switzerland, in the 1930's. After multiple revisions, the ILO system remains the most widely used method for classifying chest radiographic abnormalities related to inhalation of pathogenic dusts[23,24,25]. Up through the early 2000's, the ILO system was predicated solely on use of film-screen radiographs (FSR)[23]. Beginning in about the 1980's in the United States and other medically advanced countries, many medical facilities began replacing traditional FSR equipment with various forms of digital radiographic (DR) imaging, including both hard copy (HC) format (i.e., digital images printed on film and viewed with a traditional light box), and soft copy (SC) format (i.e., digital images viewed on a computer workstation monitor). By the early 2000's it had become difficult to obtain FSR chest radiographs in many such locations, even though FSR is still widely used in many low and middle income countries[26]. Yet, up until that point, there had only been a few studies that compared the reliability of the DR technology and FSR for the identification and quantification of abnormalities due to dust inhalation using the ILO system[27,28].

In a recently concluded study funded by the National Institutes of Occupational Safety and Health (NIOSH)[29,30], a total of one hundred seven subjects, many suffering from severe pneumoconiosis, were recruited from the pool of patients seen at or referred to the University of Michigan; or listed in the Michigan or Ohio Silicosis Registries. Pneumoconioses are a group of interstitial lung diseases caused by the inhalation of certain dusts and the lung tissue's reaction to it. The principal causes of the disease are work-place exposures, including asbestos, silica and coal.

Six certified raters evaluated images in each of the three formats (FSR, SC, and HC) in a random order, being blinded to the identity of the subject which the image was associated with. An ILO scoring sheet that included several categorical items, with ratings recorded both on nominal and ordinal scales, were filled out after each reading. This exercise was repeated once two months after the initial reading, yielding a total of six readings on each subject by a single rater. The data layout for a typical dichotomous outcome for two subjects from the study is schematically depicted in Table 1. The goal of the investigation was to assess and compare the impact of image format (FSR, SC, and HC) on the recognition and quantification of dust-related abnormalities as well as on the intra- and inter-rater reliability of the readings.

[**TABLE** 1 about here.]

3 | A BAYESIAN METHOD FOR COMPARING κ-COEFFICIENT OF DEPENDENT SAMPLES

3.1 | A Bayesian model

In order to understand the basic framework, assume each of n subjects is rated on a binary scale using m different settings. The methodology presented in this section applies to a single rater. The multiple rater problem is addressed in Section 4. The rater assigns the binary score to each subject twice using each method of the m settings. Let X_{1ij} and X_{2ij} be the ratings on the i-th subject using the j-th setting in the two rounds, respectively. For setting j, one can measure the intra-rater agreement using κ_j. Since the κ values are calculated based on data from the same set of subjects, they are correlated. We propose a Bayesian method for testing the hypothesis $H_0 : \kappa_1 = \kappa_2 = ... = \kappa_m$. Note the method proposed in this section can be applied to grouped data as shown in Figure 1 and no individual-level data are needed.

[**FIGURE** 1 about here.]

For a generic setting j, let p_{abj} denote the probability of being assigned to category a by the first reading and to category b by the second round reading ($a, b = 0, 1$) within the j-th setting ($j = 1, 2, ..., m$). The kappa coefficient for the j-th setting is

$$\kappa_j = \frac{2(p_{11j} - p_{1.j}p_{.1j})}{p_{1.j} + p_{.1j} - 2p_{1.j}p_{.1j}} \tag{2}$$

where $p_{1.j} = p_{10j} + p_{11j}$ and $p_{.1j} = p_{01j} + p_{11j}$.

Applying the multinomial distribution to the counts in the four cells in the j-th contingency table, the likelihood for j-th setting, based on a random sample of n pairs of binary responses, is

$$L_j(p_{11j}, p_{1.j}, p_{.1j}) = \frac{n!}{\prod n_{abj}!} p_{11j}^{n_{11j}} (p_{1.j} - p_{11j})^{n_{10j}} (p_{.1j} - p_{11j})^{n_{01j}} (1 - p_{1.j} - p_{.1j} + p_{11j})^{n_{00j}} \tag{3}$$

where $n_{abj} = \sum_{i=1}^{n} I(X_{1ij} = a, X_{2ij} = b)$. We assume the availability of count data which allows us to work with the likelihood in (3).

In order to formulate the dependence, one can potentially introduce a multivariate structure by specifying all joint probabilities across readings and settings. Such an overly parameterized problem may suffer from sparsity even when the number of methods m is moderate. Instead we adopt a more parsimonious structure of modeling the dependence.

Note that the marginal probabilities can be re-expressed as $p_{1.j} = p_{1.1}^{c_{j-1}}$, $p_{.1j} = p_{.11}^{c_{j-1}}$ for some positive real numbers c_{j-1}, $j = 2, ..., m$. The advantage of such a representation is the ability to induce dependence through the common baseline values $p_{1.1}, p_{.11}$. Further, the motivation of using a power formulation stems from the frailty structure often used for modeling dependence. Much like the frailty structure, c_j's are latent, but unlike frailty, they are not shared across all j. Parsimony is imposed by assuming that the same c_{j-1} operates on both marginals. This facilitates formulation of dependence between the marginals $p_{1.j}, p_{.1j}$ for the jth setting. For a dichotomous rating scale, the likelihood is given by

$$L = \prod_{j=1}^{m} L_j(p_{11j}, p_{1.j}, p_{.1j}). \tag{4}$$

For the priors, we choose the beta distribution for $p_{1.1}$ and $p_{.11}$, i.e, $Beta(\mu\alpha, \mu(1-\alpha))$, $Beta(\nu\beta, \nu(1-\beta))$, where $0 < \alpha < 1, 0 < \beta < 1$, and they are independent. The hyperparameters $\alpha = E(p_{1.1})$ and $\beta = E(p_{.11})$ reflect the average prior guess for $p_{1.1}$ and $p_{.11}$, whereas the hyperparameters μ and ν are quantification of the degree of belief about the guess. A non-informative positive prior such as $U(0, +\infty)$ is assumed for c_{j-1}. Then p_{11j} conditioned on $p_{1.j}$ and $p_{.1j}$ is $U[max(0, p_{1.j} + p_{.1j} - 1), min(p_{1.j}, p_{.1j})]$.

Based on the above prior and likelihood, when $m = 2$, the joint posterior of $p_{1.1}, p_{.11}, c_1, p_{111}, p_{112}$ is

$$\begin{aligned} \pi(p_{1.1}, p_{.11}, c_1, p_{111}, p_{112} | data) \propto\ & p_{111}^{n_{111}} (p_{1.1} - p_{111})^{n_{101}} (p_{.11} - p_{111})^{n_{011}} (1 - p_{1.1} - p_{.11} + p_{111})^{n_{001}} \\ & \times p_{112}^{n_{112}} (p_{1.1}^{c_1} - p_{112})^{n_{102}} (p_{.11}^{c_1} - p_{112})^{n_{012}} (1 - p_{1.1}^{c_1} - p_{.11}^{c_1} + p_{112})^{n_{002}} \\ \times & \frac{p_{1.1}^{\mu\alpha-1} (1 - p_{1.1})^{\mu(1-\alpha)-1} p_{.11}^{\nu\beta-1} (1 - p_{.11})^{\nu(1-\beta)-1}}{[min(p_{1.1}, p_{.11}) - max(0, p_{1.1} + p_{.11} - 1)][min(p_{1.1}^{c_1}, p_{.11}^{c_1}) - max(0, p_{1.1}^{c_1} + p_{.11}^{c_1} - 1)]} \end{aligned} \tag{5}$$

The joint distribution in (5) is simply a product of the likelihood function and the prior model, modulo the constant of integration. While (5) cannot be simplified any further, a fast and efficient Markov Chain Monte Carlo (MCMC) method can be employed in order to calculate the marginal posteriors of the parameters. The posterior distribution of κ_1, κ_2 can be obtained with (2). The posterior distribution of the difference of κ_1, κ_2 is used to test the hypothesis $H_0 : \kappa_1 = \kappa_2$. We refer to this newly proposed method as Bayesian Method for Dependent Kappa (BMDK) hereafter.

3.2 | Homogeneity of the Kappa Statistics

In general, we are interested in testing the overall hypothesis $H_0 : \kappa_1 = \cdots = \kappa_m$ vs. the alternative H_a : not H_0. In the Bayesian context, the usual method is to make a decision on the basis of Bayes Factor (BF). Below we propose this and two other simple-to-adopt methods as competitors.

Method A (Bayes Factor)

Bayes Factor (BF), originally proposed by Jeffreys[31] treats hypothesis testing as a model selection problem and is estimated as the ratio of the marginal likelihood of the data under the null and the alternative, the two competing models. Denoting by D the data evidence, we have the expression

$$BF = \frac{\Pr(D|H_0)}{\Pr(D|H_a)}. \tag{6}$$

There are several ways to estimate the quantity $\Pr(D|H)$ for a generic hypothesis H, of which perhaps the easiest, and the most easily implementable is the one that relies on evaluation of the conditional likelihood given the parameters that are sampled from the prior. Denoting by $\boldsymbol{\theta}$ the ensemble of parameters, we can write

$$\Pr(D|H) = \int \Pr(D|H,\boldsymbol{\theta})\ \Pr(\boldsymbol{\theta}|H)\ d\boldsymbol{\theta}, \tag{7}$$

where $\Pr(\boldsymbol{\theta}|H)$ refers to the prior distribution under H. One can repeatedly sample parameters $\boldsymbol{\theta}_j$ from the prior under H for a large number of times and estimate (7) by the sample average of the conditional likelihoods $\Pr(D|H,\boldsymbol{\theta}_j)$ evaluated at the sampled $\boldsymbol{\theta}_j$. We adopt this approach for our simulations provided in the next section. We follow standard guidelines for assessment of BF with BF values $\geq 1/3$ indicating insufficient evidence against H_0[32].

Method B (Probability of Dominance)

The posterior probability of dominance of the Bayesian estimators can be used as evidence either in favor or against H_0. Specifically one can compute

$$M = \max_{i \neq j} \Pr(\hat{\kappa}_i > \hat{\kappa}_j | data), \tag{8}$$

where $\hat{\kappa}$ refers to the Bayesian estimator of κ. If M is close to 50%, the evidence against H_0 is negligible. In the Bayesian paradigm, such a criterion is natural and has been used in other contexts[33]. In order to have a guideline akin to BF, one can come up with a prescription such as

M	50% – 69%	70% – 79%	80% – 89%	≥ 90%
Evidence against H_0	None	Weak	Moderate	Strong

Method C (Multivariate Statistics)

The above methods do not explicitly incorporate the extent of dependence between the estimators themselves. A properly weighted multivariate statistics can be constructed as

$$Q = \boldsymbol{\Psi}' \hat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{\Psi} \tag{9}$$

where $\boldsymbol{\Psi} = (\hat{\kappa}_2 - \hat{\kappa}_1, \hat{\kappa}_3 - \hat{\kappa}_2, \cdots, \hat{\kappa}_m - \hat{\kappa}_{m-1})'$ is the vector of the successive differences of the Bayesian estimators and $\hat{\boldsymbol{\Sigma}}$ is the corresponding estimated variance-covariance matrix. Since the inferential approach is simulation based, all the ingredients can be estimated without difficulty from the posterior MCMC samples. While the true distribution of Q under H_0 is unknown, it is reasonable to think of Q to be approximately distributed as a chi-square random variable with $(m-1)$ degrees of freedom.

3.3 | Simulation studies for evaluating BMDK

In order to assess the performance of BMDK, we ran a simulation study for $m = 3$. To evaluate Type I error of the proposed method (Table 2), summary data of 2 by 2 tables were generated for three settings from multinomial distributions determined by sample size n, fixed marginal probabilities $p_{1.1}, p_{.11}$, link parameters c_1, c_2, and a common κ. To evaluate power of the proposed method (Table 3), a similar simulation setup was used but with different κ's for the three settings. The given values of c_1, c_2 determined the marginals for the second and third settings, i.e., $p_{1.2} = p_{1.1}^{c_1}, p_{1.3} = p_{1.1}^{c_2}$.

In estimation, we consider the hyper-parameters $\alpha = \beta = 2$ and $\mu = \nu = 0.5$, which impose the prior specification, $p_{1.1}, p_{.11} \sim beta(1,1)$. Further, c_1, c_2 are independently drawn from a $U(0,10)$ distribution. Codes for the BDMK are presented in Appendix Section A.1. We fitted our model using Rjags[34]. We ran three chains and assessed convergence graphically using trace plots. We used a burn-in of 1,000 iterations and conducted inference based on a chain of length 5,000 from the posterior distributions of model parameters. Quantities in Table 2 and Table 3 are obtained as averages over 1000 replications of this process.

Table 2 shows the type I error analogs for testing $H_0 : \kappa_1 = \kappa_2 = \kappa_3$ at various levels of n, κ, and p, where κ is the common value. We estimated statistics based on the three homogeneity tests described in Section 3.2 and compared them with the Bayesian method of Basu et al[10], which is based on independence of κ statistics among settings. We report the statistics BF, M, Q as described in (6)–(9). In addition, we also report the mean squared error (MSE) estimated as $\sum_{j=1}^{3}(\hat{\kappa}_j - \kappa_j)^2/3$ where $\hat{\kappa}_j$ is the posterior mean estimate for κ for method j, $j = 1,2,3$. Based on the table values, BDMK and Basu's method are reasonably close to each other for small values of κ. Both values estimate the average BF to be close to 1 with the proportion of BF values exceeding 1/3 varying between 82% and 89%. Under either method, the dominance probability M mostly stays below 60% with BDMK based estimates falling slightly closer to 50% for larger κ. Similarly, the percentage of times Q is larger than 5.99, the 95^{th} percentile of χ^2 with 2 degrees of freedom, is greater for BDMK compared to the method by Basu et al., especially for larger κ. Finally the estimated MSE based on BDMK are also lower than its competitor.

Table 3, estimates the analog of power for detecting the difference between the κ parameters. In this case, the proportion of cases for which $BF \geq 1/3$ is significantly smaller for BDMK compared to Basu et al.'s method. This is particularly true for moderate to large n. The dominance probability M also shows a much stronger support for the alternative under BDMK. If we estimate power by $\Pr(Q > 5.99)$, again BDMK beats the Basu et al's method which ignores dependence. The average MSE estimated by $\sum_{j=1}^{3}(\hat{\kappa}_j - \kappa_j)^2/3$ are comparable across the two methods. Overall it appears that BDMK outperforms Basu et al's method with respect to power and maintains a comparable performance with regards to type I error.

Finally Figure S1 shows for each scenario in Table 2, an overlay of a smoothed empirical distribution of Q and the probability distribution function of a χ^2 with 2 degrees of freedom. The proximity of the curves justifies using χ^2 as the approximate null distribution for calibration purposes.

[**TABLE** 2 about here.]

[**TABLE** 3 about here.]

3.4 | Application of BMDK to the imaging study

To demonstrate the above proposed BMDK method using the chest radiograph imaging study, we compared the intra-rater agreement on the parenchymal abnormalities (PARABS) rating between the three chest radiographic image formats (HC, SC, FSR). Parenchymal abnormalities are characterized by abnormal increase in tissue density around the air sacs of the lung. Since some form of parenchymal abnormality is common, it is important to identify the ones which require treatment and is indicative of interstitial disease. The outcome variable (PARABS) is dichotomous, indicating the presence or absence of abnormality. A total of six experienced raters interpreted all images in each of the three image formats (FSR, HC, SC) in random order on two separate occasions (i.e., each rater read each image twice). We analyzed the data separately for these six raters. In each case, we have used the same prior specification as we used for the simulation. Table 4 reports the statistics we proposed in Section 3.2 for each rater. Instead of reporting an overall M statistic, we documented the posterior probability of the three pairwise ordering of estimated κ's, in order to identify where the difference (if any) is. For rater 3, both BF and the dominant probability method identify the ordering $\kappa_{FSR} > \max\{\kappa_{HC}, \kappa_{SC}\}$ to be present. The Q value is also quite high although did not cross the significance threshold of 5.99. No significant difference between κ_{HC} and κ_{SC} emerged. For other raters, the distinction between κ statistics is not prominent.

Figure S2 in the Appendix shows side-by-side boxplots of the reliability estimated from the MCMC samples stratified by rater, using our proposed BMDK method. As we found before, for most raters, the difference between image formats is minimal. However, the variability across raters is evident from the plots.

In our data analysis, we assumed different c values for HC vs FSR (c_1) and HC vs SC (c_2). Table S1 in the Appendix demonstrates the summary measures from the posterior stratified by raters. Despite the fact that the prior support for c is moderate, the bulk of the posterior distribution generally stayed under 3. Further, the distributions are unimodal, exhibiting only minor skewness (Figure S3). All estimates were based on 50000 MCMC samples.

Since the unadjusted calculations cannot provide an overall comparison between the three image formats adjusting for heterogeneity between raters, the conclusion is not straightforward. When individual data is available, a deeper methodology to capture subject level heterogeneity is warranted. In what follows, we present a regression based approach that can be used to adjust for subject and rater level covariates.

[**TABLE** 4 about here.]

4 | A BAYESIAN JOINT MODEL FOR COMPARING DEPENDENT κ-COEFFICIENT ADJUSTED FOR SUBJECT- AND RATER-LEVEL HETEROGENEITY

4.1 | A Bayesian joint model

Lipsitz et al.[11] proposed a regression model for κ for dichotomous ratings by using two ordinary logistic regressions and one linear regression. To take advantage of widely available standard software packages, they adopted a two-stage approach. At the first stage, one fits the two ordinary logistic regressions to calculate case-specific marginal probabilities for each pair of readings; at the second stage, these case-specific marginal probabilities are plugged into the linear regression model for κ. The caveat of this two-stage approach is that the uncertainty in the plugged-in estimates of case-specific marginal probabilities from the first stage sub-model can potentially lead to biased and less efficient estimates in the second stage sub-model[35]. We propose to eliminate this potential bias and account for hierarchical structure in agreement data by jointly estimating parameters in the two sub-models simultaneously under a Bayesian framework.

Using the same notation as in Section 3, the joint probability function for the observed pair of ratings X_{1ij} and X_{2ij} can be formulated as a function of subject-level marginal probabilities $p_{1.ij}$ and $p_{.1ij}$, and subject level Kappa κ_{ij}.

$$\begin{aligned} Pr(X_{1ij} = x_{1ij}, X_{2ij} = x_{2ij}) = {} & [p_{1.ij}p_{.1ij} + \frac{\kappa_{ij}}{2}(p_{1.ij}p'_{.1ij} + p'_{1.ij}p_{.1ij})]^{x_{1ij}x_{2ij}} \\ & \times [p_{1.ij}p'_{.1ij} - \frac{\kappa_{ij}}{2}(p_{1.ij}p'_{.1ij} + p'_{1.ij}p_{.1ij})]^{x_{1ij}(1-x_{2ij})} \\ & \times [p'_{1.ij}p_{.1ij} - \frac{\kappa_{ij}}{2}(p_{1.ij}p'_{.1ij} + p'_{1.ij}p_{.1ij})]^{(1-x_{1ij})x_{2ij}} \\ & \times [p'_{1.ij}p'_{.1ij} + \frac{\kappa_{ij}}{2}(p_{1.ij}p'_{.1ij} + p'_{1.ij}p_{.1ij})]^{(1-x_{1ij})(1-x_{2ij})} \end{aligned} \tag{10}$$

where $p'_{1.ij} = 1 - p_{1.ij}$ and $p'_{.1ij} = 1 - p_{.1ij}$, and x_{1ij}, x_{2ij} equal 0 or 1. We assume κ_{ij} is a linear function of K covariates $\boldsymbol{W}_{ij} = \{W_{1ij}, W_{2ij}, ..., W_{Kij}\}$,

$$\kappa_{ij} = \beta_0 + \boldsymbol{\beta}\boldsymbol{W}_{ij} \tag{11}$$

where $\boldsymbol{\beta} = \{\beta_1, \beta_2, ..., \beta_K\}$ are regression coefficients. We used two separate logistic regression models for modeling $p_{1.ij}$ and $p_{.1ij}$ to allow effects of subject-level covariates and rating methods on these subject-specific marginal probabilities to differ between the two rounds. Alternatively, they can be modeled using the same equation. In addition, we introduced subject level random effects to account for association between readings for the same subjects.

$$logit(p_{1.ij}) = \alpha_{1i} + \gamma_{10} + \boldsymbol{\gamma_1}\boldsymbol{Z}_{ij} \tag{12}$$

$$logit(p_{.1ij}) = \alpha_{2i} + \gamma_{20} + \boldsymbol{\gamma_2}\boldsymbol{Z}_{ij} \tag{13}$$

where $\{\alpha_{1i}, \alpha_{2i}\} \overset{i.i.d}{\sim} N(\mathbf{0}, \Sigma)$ are the subject-specific random effects, $\Sigma = \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix}$, $\boldsymbol{Z}_{ij} = \{Z_{1ij}, Z_{2ij}, ..., Z_{Sij}\}$ are S covariates that are associated with the marginal probabilities, and $\boldsymbol{\gamma_1} = \{\gamma_{11}, \gamma_{12}, ..., \gamma_{1S}\}$ and $\boldsymbol{\gamma_2} = \{\gamma_{21}, \gamma_{22}, ..., \gamma_{2S}\}$ are the corresponding regression coefficients. Combining equations (10), (11), (12) and (13), the likelihood of the sample can be written as a function of β_0, $\boldsymbol{\beta}$, γ_{10}, γ_{20}, $\boldsymbol{\gamma_1}$, $\boldsymbol{\gamma_2}$, and Σ.

$$L = \prod_{i=1}^{N}\prod_{j=1}^{m}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} Pr(X_{1ij}, X_{2ij}|\beta_0, \boldsymbol{\beta}, \gamma_{10}, \gamma_{20}, \boldsymbol{\gamma_1}, \boldsymbol{\gamma_2}, \boldsymbol{Z}_{ij}, \boldsymbol{W}_{ij}) \times f(\alpha_{1i}, \alpha_{2i}|\sigma_1^2, \sigma_2^2)d\alpha_{1i}d\alpha_{2i}$$

Denote the prior distribution for β_0, $\boldsymbol{\beta}$, γ_{10}, γ_{20}, $\boldsymbol{\gamma_1}$, $\boldsymbol{\gamma_2}$, and Σ as $\pi(\beta_0)$, $\pi(\boldsymbol{\beta})$, $\pi(\gamma_{10})$, $\pi(\gamma_{20})$, $\pi(\boldsymbol{\gamma_1})$, $\pi(\boldsymbol{\gamma_2})$, and $\pi(\Sigma)$, respectively. Denote $\boldsymbol{D}$ as observed data. MCMC can be employed to sample from the joint posterior distribution.

$$\begin{aligned} & \pi(\beta_0, \boldsymbol{\beta}, \gamma_{10}, \gamma_{20}, \boldsymbol{\gamma_1}, \boldsymbol{\gamma_2}, \Sigma|\boldsymbol{D}) \\ \propto & \prod_{i=1}^{N}\prod_{j=1}^{m}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} Pr(X_{1ij}, X_{2ij}|\beta_0, \boldsymbol{\beta}, \gamma_{10}, \gamma_{20}, \boldsymbol{\gamma_1}, \boldsymbol{\gamma_2}, \boldsymbol{Z}_{ij}, \boldsymbol{W}_{ij}) \times f(\alpha_{1i}, \alpha_{2i}|\sigma_1^2, \sigma_2^2)d\alpha_{1i}d\alpha_{2i} \\ & \times\pi(\beta_0) \times \pi(\boldsymbol{\beta}) \times \pi(\gamma_{10}) \times \pi(\gamma_{20}) \times \pi(\boldsymbol{\gamma_1}) \times \pi(\boldsymbol{\gamma_2}) \times \pi(\Sigma) \end{aligned}$$

4.2 | Application of the proposed Bayesian joint model to the imaging study

We applied the proposed Bayesian joint model to the chest radiograph imaging study. It is well known that the marginal distribution of the outcome variable affects the level of κ-coefficient and makes interpretation and comparison between κ values challenging. The goal of this analysis was to compare intra-rater agreement on the parenchymal abnormalities rating between the three chest radiographic image formats (HC, SC, FSR) adjusting for patient characteristics that can potentially influence the probability of being rated as having parenchymal abnormalities and difference in intra-rater agreement between raters. We modeled the subject-level marginal probability of being rated as having parenchymal abnormalities in round 1 and round 2 ($p_{1.ij}$ and $p_{1.ij}$) as a function of body mass index (BMI), gender, pack-years of smoking, and age of patients, and image format.

$$logit(p_{1.ij}) = \alpha_{1i} + \gamma_{10} + \gamma_{11}BMI_i + \gamma_{12}GENDER_i + \gamma_{13}PACKYEARS_i + \gamma_{14}AGE_i + \gamma_{15}HC_{ij} + \gamma_{16}FSR_{ij}$$
$$logit(p_{.1ij}) = \alpha_{2i} + \gamma_{20} + \gamma_{21}BMI_i + \gamma_{22}GENDER_i + \gamma_{23}PACKYEARS_i + \gamma_{24}AGE_i + \gamma_{25}HC_{ij} + \gamma_{26}FSR_{ij}$$

Here HC_{ij}, FSR_{ij} are indicators of ratings based on HC, FSR formats, respectively.

We modeled the κ-coefficient as a function of image format and rater.

$$\kappa_{ij} = \beta_0 + \beta_1 rater_{2ij} + \beta_2 rater_{3ij} + \beta_3 rater_{4ij} + \beta_4 rater_{5ij} + \beta_5 rater_{6ij} + \beta_6 HC_{ij} + \beta_7 FSR_{ij}$$

where $\{rater_{2ij}, ..., rater_{6ij}\}$ are five rater indicators, with rater 1 as the reference category. Using the same equations for marginal probabilities $p_{1.ij}$, $p_{.1ij}$ and κ_{ij}, we also fit a Bayesian version of the two-stage model proposed by Lipsitz et al. (2001)[11] and compared the results obtained from the joint Bayesian model to the two-stage model (Table 5, See Table S2 for more details).

We used N(0,100) as priors for β_0, $\boldsymbol{\beta}$, γ_{10}, γ_{20}, $\boldsymbol{\gamma_1}$, and $\boldsymbol{\gamma_2}$. The subject specific parameters α_{1i}, α_{2i} were assumed to be i.i.d. $N(0, \sigma^2)$ random variables, where the hyper-parameter $\sigma \sim U(0, 100)$. We fitted our model using Rjags. For generating the MCMC samples for inference, we ran three chains and assessed convergence graphically using trace plots. We used a burn-in of 1,000 iterations and conducted inference based on a chain of length 5,000 from the posterior distributions of model parameters.

The Bayesian joint model results showed that BMI, age, sex, and image format were significantly associated with the chance of being rated as having parenchymal abnormalities. After adjusting for rater difference in κ-coefficient, image format didn't significantly affect the intra-rater reliability. Although the two-stage method showed similar results, sex wasn't found to be significantly associated with the likelihood of being rated as having parenchymal abnormalities using the two-stage method.

[**TABLE** 5 about here.]

4.3 | Simulation study for evaluating the proposed Bayesian joint model

To compare the performance of joint modeling and Lipsitz's 2-stage model, we ran a simulation with a simpler setting with two imaging methods only. In the marginal models, a continuous variable z_{1i} from standard normal distribution and a binary variable z_{2i} with probability 0.5 are generated, and method indicator is included as z_{3ij}. For $\boldsymbol{Z_{ij}} = \{z_{1i}, z_{2i}, z_{3ij}\}$ in (12), we took the coefficients to be $\gamma_{10} = 1, \boldsymbol{\gamma_1} = \{1.2, -1, 0.3\}$. In marginal model for $p_{.1ij}$ in (13), $\gamma_{20} = 1.1, \boldsymbol{\gamma_2} = \{1, -0.8, 0.4\}$. The random intercept for each subject $\alpha_{1i}, \alpha_{2i} \sim N(0, 0.1^2)$, which is relatively small compared to the main effect in the models. Logistic link was used for the marginals. For the kappa model (11), only one method indicator was included. Thus when type I error is sought in Table 6, $\beta_0 = \kappa$, and $\beta = 0$. On the other hand, in Table 7 $\beta_0 = \kappa_1$ and $\beta = \kappa_2 - \kappa_1$.

We compared our Bayesian joint-modeling methods with Lipsitz's frequentist method. For the Bayesian joint model, we only report the method based on the dominant probability M. We also estimated the probability of M exceeding 90% by the corresponding empirical proportion based on the 1000 iterations. Type I error and power for the frequentist method were estimated empirically in a similar manner. The story is analogous to that observed in the case of grouped data. The Bayesian joint model outperforms the Lipsitz method with regards to power while the type I errors are generally comparable. When the random intercept variance was tripled to induce greater within-subject correlation, the pattern remained unchanged.

[**TABLE** 6 about here.]

[**TABLE** 7 about here.]

5 | CONCLUDING REMARKS

In this article, we investigated homogeneity testing of correlated kappa statistics. The study explored testing both under a grouped-data setting as well as under individual-level data setting using a regression formulation that controls for subject- and rater-level heterogeneity. The framework we adopted is Bayesian, which has the capability of handling flexible modeling structure as well as providing a natural strategy for incorporating measurement error. In general, our proposed Bayesian method has better power than the competing Bayesian or frequentist method which either ignores the intra-subject correlation, or measurement error. An additional study (not reported here) also demonstrated the superiority of BDMK over other frequentist methods that (falsely) assume interchangeability of raters (Donner et al.[13]). We proposed multiple methods for assessing homogeneity of κ statistics, all of which work well in tandem, although it appears that the dominant probability method is more pro-active in picking up signals.

The current article is entirely devoted to dichotomous ratings. With more than two categories of response, the distinction between nominal and ordinal scales comes into play, and the models and methods used in these contexts are inherently different. However, we contend that the theoretical framework of our proposed methods extend somewhat naturally. For example, in the grouped-data setting, the marginal probabilities can be modeled as a Dirichlet distribution. Analogously, c, the power parameters for each table, can also be modeled as a multivariate distribution, such as multivariate lognormal. While theoretically feasible, the challenge is the loss of parsimony and it is unclear how such modeling impacts the speed and convergence of the MCMC algorithms. In a similar token, individual-level regression models may also involve heavy parameterization. Studying agreement for polytomous response incorporating dependence is an exercise worth pursuing.

While kappa is a traditional measure that has been well accepted for decades, its image has been tainted by some shortcomings. The strong dependence of kappa on marginal prevalence, forcing an upper bound, is well known. Related to this, Feinstein and Cicchetti[36] noted that in the case of marginal heterogeneity, it is possible to get a low kappa value despite high level of diagonal agreement. Interval estimation of κ when true κ is non-zero is not straightforward. Several modifications of kappa as well as alternative measures for assessing agreement between diagnostic tests have been developed. A popular line of development envisions discrete scoring as realizations arising from an underlying (latent) scoring scheme on the continuum[37].

Application of the kappa statistic to describe agreement between two diagnostic tests with or without the presence of any gold standard can be investigated in the context of the sensitivity and specificity. Much like prevalence, κ is also dependent on these two properties of a test[38]. How this dependence affects the inference related to κ, remains an open area of research.

Agreement between diagnostic tests, rating methods etc. remains an important topic of research and exploration. Several researchers have developed methodology to assess agreement between outcomes measured on different scales. These include Concordance Correlation Coefficient introduced and studied by Lin[39,40] for continuous data that has been extended to repeated measures[41] and survival outcomes[42]. When trying to assess the association between a continuous and an ordinal measurement that have not necessarily been measured in identical manner, a more general concept of broad sense agreement has recently been proposed[43]. Bayesian treatment of such agreement paradigms have largely been unexplored.

ACKNOWLEDGEMENT

Support for this research was provided by grant #S2200-22/23 from the Association of Schools of Public Health (ASPH) and the Centers for Disease Control and Prevention (CDC)/National Institute for Occupational Safety and Health (NIOSH) and a grant from the MCUBED program, an internal funding mechanism at the University of Michigan.

References

1. Cohen J. A coefficient of agreement for nominal scales. *Educational and psychological measurement* 1960; 20(1): 37–46.
2. Cicchetti DV, Allison T. A new procedure for assessing reliability of scoring EEG sleep recordings. *American Journal of EEG Technology* 1971; 11(3): 101–110.
3. Fleiss JL, Cohen J. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and psychological measurement* 1973; 33(3): 613–619.
4. Landis JR, Koch GG. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics* 1977: 363–374.
5. Fleiss JL, Levin B, Paik MC. *Statistical methods for rates and proportions*. John Wiley & Sons . 2013.
6. Shoukri MM. *Measures of interobserver agreement and reliability*. CRC press . 2010.
7. Fleiss JL. Measuring nominal scale agreement among many raters.. *Psychological bulletin* 1971; 76(5): 378.
8. Kraemer HC. Extension of the kappa coefficient.. *Biometrics* 1980; 36(2): 207–216.
9. Donner A, Eliasziw M, Klar N. Testing the homogeneity of kappa statistics. *Biometrics* 1996: 176–183.
10. Basu S, Banerjee M, Sen A. Bayesian inference for kappa from single and multiple studies. *Biometrics* 2000; 56(2): 577–582.
11. Lipsitz SR, Williamson J, Klar N, Ibrahim J, Parzen M. A simple method for estimating a regression model for κ between a pair of raters. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 2001; 164(3): 449–465.
12. Banerjee M, Capozzoli M, McSweeney L, Sinha D. Beyond kappa: A review of interrater agreement measures. *Canadian journal of statistics* 1999; 27(1): 3–23.
13. Donner A, Shoukri MM, Klar N, Bartfay E. Testing the equality of two dependent kappa statistics. *Statistics in Medicine* 2000; 19(3): 373–387.
14. McKenzie DP, Mackinnon AJ, Péladeau N, et al. Comparing correlated kappas by resampling: is one level of agreement significantly different from another?. *Journal of psychiatric research* 1996; 30(6): 483–492.
15. Vanbelle S, Albert A. A bootstrap method for comparing correlated kappa coefficients. *Journal of Statistical Computation and Simulation* 2008; 78(11): 1009–1015.
16. Barnhart HX, Williamson JM. Weighted least-squares approach for comparing correlated kappa. *Biometrics* 2002; 58(4): 1012–1019.
17. Kang C, Qaqish B, Monaco J, Sheridan SL, Cai J. Kappa statistic for clustered dichotomous responses from physicians and patients. *Statistics in medicine* 2013; 32(21): 3700–3719.
18. Yang Z, Zhou M. Weighted kappa statistic for clustered matched-pair ordinal data. *Computational Statistics & Data Analysis* 2015; 82: 1–18.
19. Nelson KP, Edwards D. Improving the reliability of diagnostic tests in population-based agreement studies. *Statistics in medicine* 2010; 29(6): 617–626.
20. Ma Y, Tang W, Feng C, Tu XM. Inference for kappas for longitudinal study data: applications to sexual health research. *Biometrics* 2008; 64(3): 781–789.
21. Fleiss JL, Cohen J, Everitt BS. Large sample standard errors of kappa and weighted kappa. *Psychological bulletin* 1969; 72(5): 323.
22. Williamson JM, Lipsitz SR, Manatunga AK. Modeling kappa for measuring dependent categorical agreement data. *Biostatistics* 2000; 1(2): 191–202.

23. ILO . Guidelines for the Use of ILO International Classification of Radiographs of Pneumoconioses, Revised edition 2000. 2002.

24. Mulloy KB, Coultas DB, Samet JM. Use of chest radiographs in epidemiological investigations of pneumoconioses.. *Occupational and Environmental Medicine* 1993; 50(3): 273–275.

25. Pham QT. Chest radiography in the diagnosis of pneumoconiosis [Workshop Report]. *The International Journal of Tuberculosis and Lung Disease* 2001; 5(5): 478–482.

26. Franzblau A, teWaterNaude J, Sen A, et al. Comparison of digital and film chest radiography for detection and medical surveillance of silicosis in a setting with a high burden of tuberculosis. *American journal of industrial medicine* 2018; 61(3): 229–238.

27. Takashima Y, Suganuma N, Sakurazawa H, et al. A flat-panel detector digital radiography and a storage phosphor computed radiography: screening for pneumoconioses. *Journal of occupational health* 2007; 49(1): 39–45.

28. Zähringer M, Piekarski C, Saupe M, et al. Comparison of digital selenium radiography with an analog screen-film system in the diagnostic process of pneumoconiosis according to ILO classification. *RoFo: Fortschritte auf dem Gebiete der Rontgenstrahlen und der Nuklearmedizin* 2001; 173(10): 942–948.

29. Franzblau A, Kazerooni EA, Sen A, et al. Comparison of Digital Radiographs with Film Radiographs for the Classification of Pneumoconiosis1. *Academic radiology* 2009; 16(6): 669–677.

30. Sen A, Lee SY, Gillespie BW, et al. Comparing film and digital radiographs for reliability of pneumoconiosis classifications: a modeling approach. *Academic radiology* 2010; 17(4): 511–519.

31. Jeffreys H. *Theory of Probability*. Oxford:Clarendon Press. third ed. 1961.

32. Kass R, Raftery A. Bayes factor and model uncertainty. *Journal of the American Statistical Association* 1995; 90(430): 773–795.

33. Satagopan J, Sen A, Zhou Q, et al. Bayes and empirical Bayes methods for reduced rank regression models in matched case-control studies. *Biometrics* 2016: 584–595.

34. Plummer M. rjags: Bayesian graphical models using MCMC. *R package version* 2013; 3(10).

35. Carroll RJ, Ruppert D, Stefanski LA, Crainiceanu CM. *Measurement error in nonlinear models: a modern perspective*. CRC press . 2006.

36. Feinstein AR, Cicchetti DV. High agreement but low kappa: I. The problems of two paradoxes. *Journal of clinical epidemiology* 1990; 43(6): 543–549.

37. Williamson JM, Manatunga AK. Assessing interrater agreement from dependent data. *Biometrics* 1997: 707–714.

38. Feuerman M, Miller A. The kappa statistic as a function of sensitivity and specificity. *International Journal of Mathematical Education in Science and Technology* 2005; 36(5): 517–527.

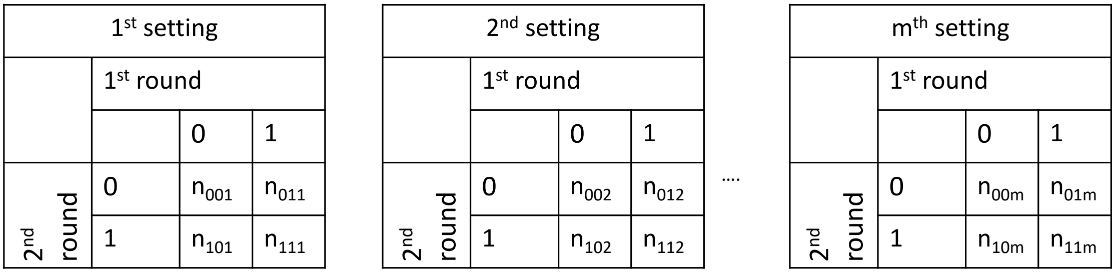
39. Lawrence I, Lin K. A concordance correlation coefficient to evaluate reproducibility. *Biometrics* 1989: 255–268.

40. Lawrence I, Lin K. Assay validation using the concordance correlation coefficient. *Biometrics* 1992: 599–604.

41. Quiroz J. Assessment of equivalence using a concordance correlation coefficient in a repeated measurements design. *Journal of Biopharmaceutical Statistics* 2005; 15(6): 913–928.

42. Guo Y, Manatunga AK. Modeling the agreement of discrete bivariate survival times using kappa coefficient. *Lifetime Data Analysis* 2005; 11(3): 309–332.

43. Peng L, Li R, Guo Y, Manatunga A. A framework for assessing broad sense agreement between ordinal and continuous measurements. *Journal of the American Statistical Association* 2011; 106(496): 1592–1601.



1st setting			
	1st round		
		0	1
2nd round	0	n_{001}	n_{011}
	1	n_{101}	n_{111}

2nd setting			
	1st round		
		0	1
2nd round	0	n_{002}	n_{012}
	1	n_{102}	n_{112}

....

mth setting			
	1st round		
		0	1
2nd round	0	n_{00m}	n_{01m}
	1	n_{10m}	n_{11m}

FIGURE 1 Summary data of dependent binary agreement data

TABLE 1 Partial Data from two selected subjects from the NIOSH study

Subject	rater	Setting					
		FSR (j=1)		HC (j=2)		SC (j=3)	
		Round 1	Round 2	Round 1	Round 2	Round 1	Round 2
i		X_{1i1}	X_{2i1}	X_{1i2}	X_{2i2}	X_{1i3}	X_{2i3}
1001	1	1	1	1	1	1	1
1001	2	1	1	1	1	1	1
1001	3	0	0	1	1	1	1
1001	4	1	1	1	1	1	1
1001	5	0	1	1	1	1	1
1001	6	1	1	1	1	1	1
1002	1	1	1	0	1	1	1
1002	2	1	1	1	1	1	1
1002	3	1	1	1	1	1	1
1002	4	0	1	1	1	0	1
1002	5	1	1	1	1	1	1
1002	6	1	1	1	1	1	1

TABLE 2 Type I errors for testing $H_0 : \kappa_1 = \kappa_2 = \kappa_3$. comparison of the proposed Bayesian method and existing method. Total of 1000 iterations. $c_1 = 0.1, c_2 = 0.25$.

	BDMK					Basu				
sample size	avg BF	BF≥1/3 (%)	M	$Q > 5.99$ (%)	MSE	avg BF	BF≥1/3 (%)	M	$Q > 5.99$ (%)	MSE
κ=0.1, (p1.1,p.11)=(0.2, 0.7)										
50	1.018	0.852	0.610	0.010	0.006	0.995	0.849	0.577	0.015	0.008
100	1.109	0.850	0.578	0.034	0.004	1.076	0.859	0.552	0.026	0.005
200	1.220	0.857	0.552	0.044	0.002	1.220	0.880	0.529	0.029	0.003
κ=0.2, (p1.1,p.11)=(0.3, 0.7)										
50	0.999	0.849	0.559	0.014	0.009	0.933	0.828	0.526	0.022	0.012
100	1.058	0.821	0.536	0.039	0.006	1.045	0.827	0.505	0.033	0.008
200	1.192	0.849	0.524	0.062	0.004	1.175	0.858	0.508	0.063	0.005
κ=0.4, (p1.1,p.11)=(0.4, 0.6)										
50	0.992	0.816	0.512	0.028	0.017	1.008	0.814	0.549	0.064	0.021
100	1.125	0.848	0.510	0.058	0.011	1.161	0.835	0.535	0.071	0.013
200	1.273	0.851	0.511	0.051	0.006	1.270	0.867	0.536	0.058	0.007
κ=0.6, (p1.1,p.11)=(0.6, 0.7)										
50	1.069	0.876	0.562	0.017	0.024	1.062	0.844	0.647	0.068	0.037
100	1.105	0.845	0.552	0.032	0.016	1.186	0.836	0.621	0.086	0.022
200	1.148	0.849	0.533	0.048	0.009	1.256	0.849	0.592	0.071	0.012

TABLE 3 Empirical powers for testing $H_0 : \kappa_1 = \kappa_2 = \kappa_3$, comparison of the proposed Bayesian estimation and existing method. Total of 1000 iterations. $c_1 = 0.1, c_2 = 0.25$.

	BDMK					Basu				
sample size	avg BF	BF≥1/3 (%)	M	$Q > 5.99$ (%)	MSE	avg BF	BF≥1/3 (%)	M	$Q > 5.99$ (%)	MSE
κ_1=0.2, κ_2=0.3, κ_3=0.4, (p1.1,p.11)=(0.3, 0.7)										
50	0.594	0.602	0.853	0.042	0.033	0.637	0.613	0.765	0.122	0.041
100	0.256	0.242	0.949	0.390	0.033	0.531	0.483	0.858	0.249	0.038
200	0.050	0.033	0.991	0.894	0.034	0.316	0.290	0.937	0.512	0.037
κ_1=0.4, κ_2=0.5, κ_3=0.6, (p1.1,p.11)=(0.4, 0.6)										
50	0.757	0.684	0.785	0.069	0.071	0.840	0.712	0.719	0.123	0.074
100	0.510	0.471	0.884	0.278	0.073	0.668	0.564	0.827	0.225	0.075
200	0.232	0.203	0.961	0.628	0.074	0.445	0.397	0.918	0.404	0.074
κ_1=0.2, κ_2=0.4, κ_3=0.6, (p1.1,p.11)=(0.4, 0.6)										
50	0.332	0.323	0.928	0.379	0.086	0.413	0.398	0.892	0.345	0.087
100	0.107	0.088	0.983	0.795	0.087	0.175	0.147	0.967	0.662	0.088
200	0.009	0.004	0.999	0.988	0.088	0.028	0.020	0.996	0.947	0.088

TABLE 4 Intra-rater comparison of kappa's for PARABS

			Posterior Probability			Posterior mean of κ		
Raters	BF	Q	$\kappa_{HC} < \kappa_{FSR}$	$\kappa_{HC} < \kappa_{SC}$	$\kappa_{FSR} < \kappa_{SC}$	HC	FSR	SC
1	0.814	1.663	0.13	0.114	0.463	0.757	0.691	0.684
2	1.959	0.307	0.437	0.645	0.715	0.697	0.685	0.739
3	0.255	4.648	0.954	0.428	0.03	0.617	0.791	0.593
4	2.559	0.157	0.527	0.651	0.629	0.745	0.75	0.779
5	1.506	0.943	0.243	0.183	0.412	0.755	0.68	0.654
6	1.453	0.980	0.668	0.838	0.723	0.758	0.796	0.842

TABLE 5 Comparison of kappa's for PARABS using a Bayesian joint model and a two-stage model

Method		Posterior Mean	Posterior SD	2.5% quantile	97.5% quantile	Posterior probability of parameter <0
Two-stage	HC vs. SC: β_6	0.101	0.091	-0.080	0.278	0.127
	FSR vs. SC: β_7	0.050	0.091	-0.125	0.233	0.288
	FSR vs. HC: $\beta_7 - \beta_6$	-0.051	0.089	-0.227	0.125	0.718
Joint model	HC vs. SC: β_6	0.061	0.041	-0.012	0.144	0.062
	FSR vs. SC: β_7	0.042	0.047	-0.051	0.130	0.193
	FSR vs. HC: $\beta_7 - \beta_6$	-0.019	0.032	-0.086	0.039	0.728

TABLE 6 Type I errors for testing $H_0 : \kappa_1 = \kappa_2$. comparison of the Bayesian joint-modeling and Lipsitz's frequentist method. Total of 1000 iterations.

Joint model				Lipsitz
κ	sample size	M	$\Pr(M > 0.9)$	type 1 error (Wald test)
random intercept $\sim N(0, 0.1^2)$				
0.5	50	0.493	0.046	0.057
	100	0.471	0.056	0.049
	200	0.474	0.076	0.067
0.3	50	0.486	0.038	0.061
	100	0.478	0.053	0.042
	200	0.500	0.066	0.052
0.1	50	0.496	0.017	0.058
	100	0.502	0.029	0.040
	200	0.502	0.050	0.032
random intercept $\sim N(0, 0.3^2)$				
0.3	50	0.486	0.031	0.053
	100	0.487	0.065	0.055
	200	0.478	0.070	0.040
0.1	50	0.478	0.011	0.048
	100	0.503	0.032	0.050
	200	0.508	0.037	0.041

TABLE 7 Powers for testing $H_0 : \kappa_1 = \kappa_2$. comparison of the Bayesian joint-modeling and Lipsitz's frequentist method. Total of 1000 iterations.

			Joint model		Lipsitz
κ_1	κ_2	sample size	M	$\Pr(M > 0.9)$	power (Wald test)
random intercept $\sim N(0, 0.1^2)$					
0.3	0.5	50	0.694	0.218	0.169
		100	0.794	0.407	0.257
		200	0.891	0.647	0.442
0.1	0.5	50	0.842	0.467	0.439
		100	0.951	0.838	0.686
		200	0.994	0.992	0.943
0.1	0.3	50	0.668	0.117	0.116
		100	0.777	0.331	0.200
		200	0.894	0.664	0.392
0.1	0.4	50	0.766	0.307	0.264
		100	0.882	0.610	0.417
		200	0.967	0.910	0.713
random intercept $\sim N(0, 0.3^2)$					
0.1	0.3	50	0.678	0.140	0.141
		100	0.770	0.342	0.228
		200	0.888	0.637	0.368